(2)

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 89-0914
6a. NAME OF PERFORMING ORGANIZATION George Mason University	6b. OFFICE SYMBOL (If applicable)
6c. ADDRESS (City, State, and ZIP Code) 4400 University Drive Fairfax VA 22030	7a. NAME OF MONITORING ORGANIZATION Air Force Office of Scientific Research
	7b. ADDRESS (City, State, and ZIP Code) Building 410 Bolling AFB, DC 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION AFOSR	8b. OFFICE SYMBOL (If applicable) NM
8c. ADDRESS (City, State, and ZIP Code) Building 410 Bolling AFB, DC 20332-6448	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER AFOSR-87-0179
	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 61102F; PROJECT NO. 2304; TASK NO. A5; WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
Hyperdimensional Data Analysis and Statistical Inference

12. PERSONAL AUTHOR(S)
Professor Edward J Wegman

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM 1 Apr 87 TO 3 Mar 89	May 18, 1989	

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This research project was based on belief that modern technology has substantially changed the flavor of problems being presented to the statistician. Electronic instrumentation implies an ability to acquire a large amount of high dimensional data very rapidly. While such capabilities have existed for some time, the emergence of cheap RAM in the 1980's has given us the ability to store and access that data in an active computer memory. ~~We contend that~~ this represents a challenge for statisticians which is substantially different in kind. The majority of existing methodology is focused on the univariate, iid random variable model. Even in the circumstance that a multivariate model is allowed, it is usually assumed to be multivariate normal. ~~We contend, in addition, that~~ while arbitrary sample size is frequently assumed, the truth of the matter is that these techniques implicitly assume small to moderate sample sizes. For example, a regression problem with 5 design variables and 1000 observations would represent no problem for traditional techniques. By contract, a regression problem with 40,000 design variables and 8 million observations would. The reason is clear. In the former case the emphasis is on statistical efficiency which is the operational goal for most current statistical

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL Lt Col Jon Berkeley	22b. TELEPHONE (Include Area Code) (202) 767-4940
	22c. OFFICE SYMBOL NM

DD Form 1473, JUN 86 — Previous editions are obsolete. — SECURITY CLASSIFICATION OF THIS PAGE

technology. By contrast, in the latter case, emphasis must be clearly on computational efficiency The emphasis on parsimony in many contemporary books and papers is a further reflection of the mind-set that implicitly focuses on small to moderate sample sizes since few parameters do not make sense in the context of very large sample sizes. Finally, we note that the very fact of largeness in sample size implies that it is unlikely we would see iid homogeneity.

Final Technical Report

Hyperdimensional Data Analysis and Structural Inference

8/6/89

Air Force Office of Scientific Research

Grant AFOSR-87-0179

PI: Professor Edward J. Wegman

Center for Computational Statistics

George Mason University

Fairfax, VA 22030

May 18, 1989

This research project was based on our belief that modern technology has substantially changed the flavor of problems being presented to the statistician. Electronic instrumentation implies an ability to acquire a large amount of high dimensional data very rapidly. While such capabilities have existed for some time, the emergence of cheap RAM in the 1980's has given us the ability to store and access that data in an active computer memory. We contend that this represents a challenge for statisticians which is substantially different in kind. The majority of existing methodology is focused on the univariate, iid random variable model. Even in the circumstance that a multivariate model is allowed, it is usually assumed to be multivariate normal. We contend, in addition, that while arbitrary sample size is frequently assumed, the truth of the matter is that these techniques implicitly assume small to moderate sample sizes. For example, a regression problem with 5 design variables and 1000 observations would represent no problem for traditional techniques. By contrast, a regression problem with 40,000 design variables and 8 million observations would. The reason is clear. In the former case the emphasis is on statistical efficiency which is the operational goal for most current statistical technology. By contrast, in the latter case, emphasis must be clearly on computational efficiency. The emphasis on parsimony in many contemporary books and papers is a further reflection of the mind-set that implicitly focuses on small to moderate sample sizes since few parameters do not make sense in the context of very large sample sizes. Finally, we note that the very fact of largeness in sample size implies that it is unlikely we would see iid homogeneity.

Thus, a new perspective is required for large, high dimensional data sets. This project was intended to explore a variety of mathematics and statistical theory related to this perspective. Specifically, we focused on graphical methods for data representation in higher dimensions, structural inference and the computational issues related to these problems. We note that in a highly multivariate setting, there is increased opportunity to focus on the functional relationship between random variables (what we refer to as structural inference) rather than simply focusing on the probability distribution of random variable (traditional statistical inference). This perspective has been discussed in more detail in papers 12 and 17 listed below. All of these topics we believe will intertwine to form the conceptual core of our approach to this new perspective. The tasks we proposed focus respectively on graphical issues (item 1), on structural inference issues (item 2) and on computational issues (item 3).

The items listed below are publications produced under this AFOSR grant during the last two years.

BOOKS OR SPECIAL ISSUES OF JOURNALS

1. ***Brain Structure, Learning and Memory,*** edited with Joel L. Davis and Robert W. Newburgh, Westview Press, Inc.: Boulder, CO for AAAS: Washington, DC, 1988

2. ***Topics in Non-Gaussian Signal Processing,*** edited with Stuart C. Schwartz and John B. Thomas, Springer-Verlag: New York, 1988

3. ***Assessing Uncertainty,*** Special Issue of the Journal of Statistical Planning and Inference, guest edited with H. Solomon, Vol. 20, No. 3, 1988

4. ***Proceedings of the 20th Symposium on the Interface: Computing Science and Statistics,*** edited with Donald T. Gantz and John J. Miller, American Statistical Association: Washington, DC, 1989

PAPERS

5. "Computational relevance of the Bayesian paradigm," Ann. Ops. Research, 9, 629-633, 1987

6. "Vector function estimation using splines," with J. Miller, J. Statist. Plan. Infer., 17, 173-180, 1987

7. "A parallel coordinate approach to statistical graphics," National Computer Graphics Association Conference Proceedings, 3, 574-580, 1987

8. "Commentary on defense funding," Notices Am. Math. Soc., 34(4), 616-618, 1987

9. "Reproducing kernel Hilbert spaces," Encyclopedia of Statistical Sciences, (N. Johnson, S. Kotz and C. Read, eds.), 8, 81-84, John Wiley and Sons: New York, 1988

10. "Sobolev spaces," Encyclopedia of Statistical Sciences, (N. Johnson, S. Kotz and C. Read, eds.), 8, 535-537, John Wiley and Sons: New York, 1988

11. "Statistical software," with Annie Hayes, Encyclopedia of Statistical Sciences, (N. Johnson, S. Kotz and C. Read, eds.), 8, 667-674, John Wiley and Sons: New York, 1988

12. "A view of computational statistics and its curriculum," Am. Statist. Assoc. Proc. Sect. Statist. Educat. 1-6, 1988

13. "Invited discussion: Application of recent methodology in statistical graphics for nonspecialists," Am. Statist. Assoc. Proc. Sect. Statist. Graphics, p. 48, 1988

14. "Introduction to assessing uncertainty," with H. Solomon, J. Statist. Planning Infer., 20, 241-244, 1988

15. "On randomness, determinism and computability," J. Statist. Planning Infer., 20, 279-294, 1988

16. "A graphical tool for distribution and correlation analysis of multiple time series," with C. Shull, in Topics in Non-Gaussian Signal Processing, (E. Wegman, S. Schwartz and J. Thomas, eds.), Springer-Verlag: New York, 1988.

17. "Computational statistics: a new agenda for statistical theory and practice," J. Washington Academy of Science, 1989.

18. "On some graphical representations of multivariate data," with Masood Bolorforoush, Proceedings of the 20th Interface Symposium on Computing Science and Statistics, 1989.

19. "Stochastic load balancing in parallel computers," to appear Proceedings of 4th Conference on Hypercube Concurrent Computers and Applications, 1989.

TECHNICAL REPORTS

20. "Transient signal detection and estimation: a survey," with Hung T. Le, Center for Computational Statistics Technical Report No. 31, George Mason University, 1988.

21. "Underwater transient signal characterization," with Hung T. Le, Center for Computational Statistics Technical Report No. 35, George Mason University, 1988.

22. "Parallelizing multiple linear regression for speed and redundancy: an empirical study," with John J. Miller and Mingxian Xu, Center for Computational Statistics Technical Report No. 39, George Mason University, 1989.

23. "Optimal estimation of transient signals via recursive splines," with Hung T. Le, Center for Computational Statistics Technical Report No. 40, George Mason University, 1989.

24. "Parallel computing and statistics," Center for Computational Statistics Technical Report No. 41, George Mason University, 1989. (Special Invited Paper for the ASA Sesquicentennial.)

25. Hung Tri Le, "A Functional Analytic Approach to Transient Signal Detection and Estimation," Center of Computational Statistics Technical Report 45, George Mason University, 1989

SOFTWARE

26. ***Mason Hypergraphics***, copyright (c) 1988, 1989 by Edward J. Wegman, a MS-DOS package for high-dimensional data analysis.

Items 7, 13, 16, 18 and 26 address graphical issues and detail a series of graphical devices we have invented to assist in the development of high-dimensional data analysis. Items 5, 11, 14, 19, 22 and 24 have a primarily computational flavor and focus on improving algorithms. Items 6, 9, 10, 20, 21 and 23 focus on structural inference issues. Items 8, 12, 15 and 17 are contibutions which are general in nature.

In addition to the technical reports and papers, the project has produced a Ph.D. student, Hung Tri Le, who completed his dissertation on the topic, A Functional Analytic Approach to Transient Signal Detection and Estimation. In addition, we have had to Masters students, Masood Bolorforoush and Mingxian Xu. We believe that the project has been extraordinarily productive. We are very grateful for the opportunity to work under Air Force sponsorship. It has allowed us not only to develop a substantial amount of research, but also to to develop our facilties and the technical skills of some of our younger faculty and students. The related instrumentation funding has allowed us to install a Intel iPSC/2 d4/VX message passing parallel computer as well as a Silicon Graphics IRIS 4D-120-GTB Graphics Workstation. Both of these facilities as well as the faculty development will pay

dividends for the Air Force and for the general scientific community well beyond the term of the contract.